Comparison of orthologous and paralogous DNA flanking the wheat high molecular weight glutenin genes: sequence conservation and divergence, transposon distribution, and matrix-attachment regions

O.D. Anderson, L. Larka, M.J. Christoffers, K.F. McCue, and J.P. Gustafson

Abstract: Extended flanking DNA sequences were characterized for five members of the wheat high molecular weight (HMW) glutenin gene family to understand more of the structure, control, and evolution of these genes. Analysis revealed more sequence conservation among orthologous regions than between paralogous regions, with differences mainly owing to transposition events involving putative retrotransposons and several miniature inverted transposable elements (MITEs). Both *gyspy*-like long terminal repeat (LTR) and non-LTR retrotransposon sequences are represented in the flanking DNAs. One of the MITEs is a novel class, but another MITE is related to the maize *Stowaway* family and is widely represented in Triticeae express sequence tags (ESTs). Flanking DNA of the longest sequence, a 20 425-bp fragment including and surrounding the HMW-glutenin *Bx7* gene, showed additional cereal gene-like sequences both immediately 5′ and 3′ to the HMW-glutenin coding region. The transcriptional activities of sequences related to these flanking putative genes and the retrotransposon-related regions were indicated by matches to wheat and other Triticeae ESTs. Predictive analysis of matrix-attachment regions (MARs) of the HMW glutenin and several α-, γ-, and ω-gliadin flanking DNAs indicate potential MARs immediately flanking each of the genes. Matrix binding activity in the predicted regions was confirmed for two of the HMW-glutenin genes.

Key words: wheat, glutenins, evolution, matrix-attachment regions, transposable elements.

Résumé : De grandes régions d'ADN adjacentes aux cinq gènes codant pour des gluténines de poids moléculaire élevé (HMW) ont été caractérisées afin de mieux connaître la structure, la régulation et l'évolution de ces gènes. L'analyse a révélé davantage de conservation de séquences parmi les régions orthologues qu'entre les régions paralogues, les différences étant principalement attribuables à des événements de transposition impliquant des rétrotransposons putatifs et plusieurs MITE (« miniature inverted transposable elements »). Tant des rétrotransposons à LTR (« long terminal repeats ») de type *gypsy* que des rétrotransposons sans LTR ont été trouvés dans les séquences adjacentes. Un des MITE est d'un type nouveau tandis qu'un autre est apparenté à la famille des éléments *Stowaway* et se trouve dans plusieurs EST (« express sequence tags ») provenant de hordées. L'ADN adjacent au gène parmi le plus grand segment séquencé, un fragment de 20 425-pb incluant et bordant le gène codant pour la gluténine-HMW *Bx7*, comprenait des séquences ressemblant à des gènes de graminées. Une activité transcriptionnelle de séquences homologues à ces gènes putatifs ou aux régions apparentées à des rétrotransposons est suggérée du fait que de telles séquences ont été trouvées parmi des collections d'EST du blé et d'autres graminées. Une analyse visant à identifier des régions d'attachement à la matrice (MAR) a suggéré la présence de possibles régions MAR dans le voisinage immédiat de tous les gènes codant pour des gluténines-HMW et de plusieurs gliadines α, γ et ω. Une activité d'attachement à la matrice a été confirmée pour deux des régions MAR prédites au voisinage de gènes codant pour des gluténines-HMW.

Mots clés : blé, gluténines, évolution, régions d'attachement à la matrice, éléments transposables.

[Traduit par la Rédaction]

Received 16 July 2001. Accepted 19 October 2001. Published on the NRC Research Press Web site at http://genome.nrc.ca on 1 March 2002.

Corresponding Editor: G.J. Scoles.

O.D. Anderson[1] and K.F. McCue. Western Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 800 Buchanan Street, Albany, CA 94710, U.S.A.
L. Larka. Lawrence Berkeley National Laboratories, 1 Cyclotron Road, Berkeley, CA 94720, U.S.A.
M.J. Christoffers. Department of Plant Sciences, North Dakota State University, Fargo, ND 58105, U.S.A.
J.P. Gustafson. USDA–ARS, Department of Agronomy, University of Missouri, Columbia, MO 65211, U.S.A.

[1] Corresponding author (e-mail: oanderson@pw.usda.gov).

DOI: 10.1139/G01-137

Introduction

The wheat (*Triticum aestivum* L. em Thell) high molecular weight (HMW) glutenin genes are the most studied wheat genes because they encode the wheat gluten protein subunits critical to the physical and chemical properties of wheat doughs (Shewry et al. 1996). The HMW-glutenin genes have also been a major focus of efforts in wheat bioengineering (Vasil and Anderson 1997). These genes are encoded at the orthologous *Glu-1* loci on each of the long arms of the three homoeologous group 1 chromosomes of hexaploid bread wheats. Each locus contains two paralogous HMW-glutenin genes encoding x- and y-type HMW-glutenin subunits (Shewry et al. 1992).

Studies have shown that wheat quality is related to differences in the quantity of HMW glutenin expressed, the genome of origin of orthologous HMW-glutenin genes, and alleles of several specific genes (Payne 1987; Macritchie 1992). Although both the HMW-glutenin genes and the proteins they encode have been studied extensively, research on the genes has focused on coding and proximal promoter sequences (Shewry et al. 1992; Anderson et al. 1998). Nothing is yet known of the more distal DNA sequences or chromosomal organization of the *Glu-1* loci and such information is relevant for a more complete understanding of the control of the HMW-glutenin genes. Individual protein levels vary among cultivars (Marchylo et al. 1992; Kolster et al. 1993), but no apparent explanation is available from sequence data variations (Anderson et al. 1998). A related observation, whose explanation may lie in some of the more distal flanking DNA sequences, comes from transformation experiments. It has been shown that transformation with native or modified HMW-glutenin genes controlled by the immediate flanking DNA shows high levels of stable expression (Blechl and Anderson 1996), in contrast to the usual case with transgenes introduced into plants (Depicker and van Montagu 1997). These wheat transformation experiments included several thousand base pairs of 5′ and 3′ flanking sequence that may contain elements that buffer expression variation. One candidate element could be matrix attachment regions (MARs) that anchor chromatin to a nuclear protein scaffold and contribute to the organization of active chromatin domains (Allen et al. 1993; Spiker and Thompson 1996; Van der Geest and Hall 1997).

In addition to the importance of understanding more about the DNAs flanking the HMW-glutenin genes, relatively little is known of the general organization of the wheat genome. The closest relevant information are the sequence of a 60-kb portion of the barley (*Hordeum vulgare* L.) genome (Panstruga et al. 1998) and a 66-kb contiguous region around the barley *mlo* gene (Shirasu et al. 2000). These results indicate a much closer spacing of genes than expected for the size of the barley genome and supports the concept that the Triticeae genomes are organized into gene-rich islands separated by non-coding regions containing retrotransposons (Feuillet and Keller 1999). No similarly extended sequence is yet reported for wheat.

The present study reports on the extended flanking sequences for five of the ‘Cheyenne’ HMW-glutenin genes, including a 20-kb contiguous sequence containing the x-type HMW-glutenin gene from the 1B chromosome (the *Bx7* gene). Features of the sequences include transposon clusters, limits of paralogous gene sequence conservation, potential early steps in orthologous region divergence, additional genes closely linked to the HMW glutenins, and the detection of MAR elements adjacent to all HMW-glutenin genes plus several other classes of endosperm-specific wheat storage protein genes.

Materials and methods

HMW-glutenin clone isolation and sequencing

The wheat HMW-glutenin genes are found at the compound *Glu-1* loci on each of the three homoeologous chromosomes of bread wheats. Each locus consists of the x- and y-type genes *Glu-1-1* and *Glu-1-2*, respectively. A further designation signifies the genome origin (A, B, and D for hexaploid bread wheats). Thus, *Glu-A1-1* is the x-type gene from the A-genome (Table 1). Common usage also includes a number designation for each HMW-glutenin protein and a shorthand common name for the gene and protein; i.e., Table 1 describes the complete set of six HMW-glutenin genes and clones from the wheat ‘Cheyenne’. The *Ay* gene in ‘Cheyenne’ is silent. The common names of the genes (*Ax2**, *Ay*, *Bx7*, *By9*, *Dx5*, *Dy10*), as well as the proteins that they encode, will be used in this paper.

All six HMW-glutenin genes from ‘Cheyenne’ have been cloned as *Eco*RI fragments, and partial sequences reported previously (Table 1). Clones used in the present study were the same as previously reported for three HMW-glutenin genes: *Ax2** (Anderson and Greene 1989), *Bx7* (Anderson and Greene 1989), and *Dy10* (Anderson et al. 1989). Other gene sequences are taken from previous publications: α-gliadins (Anderson et al. 1997), γ-gliadins (Anderson et al. 2001), and an ω-gliadin (Hsia and Anderson 2001). New individual clone isolates were used for the *Ay* and *Dx5* genes. The complete sequences of five of the gene-containing *Eco*RI fragments (*Ax2**, *Ay*, *Bx7*, *Dx5*, *Dy10*) were accomplished by walking with oligonucleotide primers in both directions over the *Eco*RI fragment or similarly sequencing *Hin*dIII and *Hin*dIII + *Eco*RI fragments of the Bx7 clone. Confirmation of the correct Bx7 subclone order was by sequencing across fragment boundaries using the intact λ clone.

Sequence analysis was performed using the Lasergene package of analysis modules (DNAStar Inc., Madison, Wis.), and Internet resources at the National Center for Biological Information (BLAST; www.ncbi.gov), the National Center for Genomic Information (MarFinder; www.ncgi.org; Kramer et al. 1996; Singh et al. 1997), and the University of Virginia (FASTA; http://fasta.bioch.virginia.edu/fasta/cgi/searchx.cgi?pgm = fa).

Matches to known sequences, either by the BLAST or FASTA algorithms, are reported with the GenBank accession number of the match and the probability (“e-value”) of the match being by chance. Generally, probabilities higher than between e^{-5} and e^{-10} (depending on the criteria being used and the researchers predilections) are considered too high to be reliably significant. In the present report, some matches with higher probabilities are reported where the sequence is the best match to known sequences and is from a species in the Triticeae tribe, such as wheat, barley, or rye (*Secale cereale* L.).

Table 1. Nomenclature, clones, and sequence data of the HMW-glutenin genes.

Gene[a]	Subunit[b]	Clone[c]	Reference[d]	Previous length[e]	Current length[f]	GenBank accession[g]
Glu-A1-1	Ax2*	λ1B1–2; pK-Ax2E	Anderson and Greene (1989) Anderson et al. (1998)	3836	6837	M22208
Glu-A1-2	Ay	λHMW50; pK+AyE11	Forde et al. (1985)	2915	8119	X03042
Glu-B1-1	Bx7	λR23; multiple[h]	Anderson and Greene (1989) Anderson et al. (1998)	4034	20425	X13927
Glu-B1-2	By9	λHMW47	Halford et al. (1987)	2996	2996	X61026
Glu-D1-1	Dx5	λ8B-1; pK+Dx5B	Anderson et al. (1989)	3590	8435	X12928
Glu-D1-2	Dy10	λ1A-2; pK-Dy10A	Anderson et al. (1989) Anderson et al. (1998)	3528	6461	X12929

[a] The names of the two genes within the compound *Glu-1* loci include a letter indicating the genome and a -1 or -2 for the x- and y-type HMW-glutenin genes, respectively.

[b] Common names of HMW-glutenin proteins and genes found in wheat cultivar Cheyenne.

[c] λ clone names and plasmid subclones of *Eco*RI fragments containing 'Cheyenne' HMW-glutenin genes.

[d] References for previously reported gene sequences.

[e] Total previously reported DNA sequence length around and including the HMW-glutenin genes of 'Cheyenne'.

[f] Total known DNA sequence length of the *Eco*RI fragments containing the HMW-glutenin genes in 'Cheyenne'. The *By9* sequence is the originally reported sequence and not the entire *Eco*RI fragment.

[g] Genbank accession No. of total HMW-glutenin fragment DNA sequences.

[h] The original λ insert was subcloned into a complete set of *Hin*dIII and *Hin*dIII + *Eco*RI subclones into pBluescriptKS-.

Nuclear matrix assay

Nuclear matrix assays were conducted using 'Columbus' wheat germ obtained from A. K. Sarkar (Canadian International Grains Institute, Winnipeg, Man.). Nuclear isolation was based on the protocol of Spiker et al. (1983). All steps in the isolation of nuclei were performed at 4°C. Five grams of wheat germ were stirred in a hexylene–glycol buffer containing Triton X-100 (Sigma, St. Louis, Mo.) and homogenized with an Ultra-Turrax polytron (IKA Works Inc., Staufen, Germany). The homogenization was filtered through a series of nylon meshes of decreasing pore size (Tetko Inc., Briarcliff Mannor, N.Y.) and nuclei were purified on a discontinuous Percoll gradient (Sigma). Nuclei banded at the 30–60% Percoll interface and were subsequently washed three times with hexylene–glycol buffer not containing Triton X-100. A 5-μL sample of nuclei was diluted to 500 μL with 5.5 M urea and 2.2 M NaCl, and used to quantify the nuclear yield via spectrophotometry at 260 nm (Hall and Spiker 1994). The remaining nuclear preparation was centrifuged, resuspended in hexylene–glycol buffer with 50% glycerol and no Triton X-100 to a concentration of 10 A_{260} U/mL, and stored at –80°C.

Nuclear matrix preparations and MAR binding assays were based on the protocol of Hall and Spiker (1994). Nuclear matrices were first prepared by thawing 10 A_{260} mL units (1 mL) of wheat germ nuclei (roughly 25 million) on ice, followed by the addition of $CuSO_4$ to a concentration of 1 mM, and incubation at 42°C for 10–15 min. Histone proteins were removed using a lithium diiodosalicylate (LIS) extraction method (Mirkovitch et al. 1984) with 10 mM LIS. After additional washes, the DNA that attached to the nuclear matrices was cleaved with a *Bam*HI–*Hin*dIII double digest for assays involving the *Bam*HI–*Hin*dIII-digested plasmid pK+Dx5B (containing the HMW-glutenin *Dx5* gene), and an *Eco*RI–*Xho*I double digest for assays involving *Pst*I–*Xho*I-digested pK-Dy10A (containing the *Dy10* gene).

DNA of the plasmid clone pK+Dx5B was prepared for assay by a double digestion with *Bam*HI and *Hin*dIII. DNA of the plasmid clone pK-Dy10A was double digested with *Pst*I and *Xho*I. Exogenous binding assays were performed by incubating 50 ng of digested pK+Dx5B or pK-Dy10A with matrices equivalent to 2 A_{260} mL units (roughly 5 million) nuclei followed by centrifugation. Matrix-bound DNA was pelleted, whereas unbound DNA was collected as supernatant. Equal pellet and supernatant fractions equivalent to 2.5 ng of pK+Dx5B or pK-Dy10A input DNA were run on agarose gels and analyzed by Southern hybridization. Digested pK+Dx5B or pK-Dy10A plasmid DNA samples were used as probes in the Southern analysis. Hybridization to endogenous wheat germ DNA did not obscure bands corresponding to analyzed plasmid input DNA owing to the relatively high abundance of the latter.

Fragments were labeled with $[^{32}P]$dCTP using High Prime random-Primer DNA labeling mix from Boehringer Mannheim, Indianapolis, Ind. (now Roche Molecular Biochemicals).

Results

Sequencing additional flanking DNA to HMW-glutenin genes

Previous reports on the HMW-glutenin genes have described coding and immediate flanking DNA sequences from the hexaploid wheat 'Cheyenne' (Table 1). We have now extended the known sequences of five of these six cloned HMW-glutenin genes. This includes the entire cloned *Eco*RI DNA fragments as shown in Table 1 from the three orthologous x-type HMW-glutenin genes (8.4 kb for the *Ax2** fragment, 6.8 for the *Bx7* fragment, and 20.4 for the *Dx5* fragment) and two of the three orthologous y-type genes (8.2 kb for the *Ay* fragment and 6.4 for the *Dy10* fragment). The *By9* gene was not included in this study. The additional sequence length ranges from almost 3000 bp for *Dy10* to over 16 000 bp for *Bx7*, and the total known sequences for this set of orthologous and paralogous genes now totals 53 273 bp.

Fig. 1. Dot plot analysis of x-type orthologous HMW-glutenin fragments. The sequence of all three ‘Cheyenne’ x-type orthologous HMW-glutenin gene fragments were compared in all possible combinations as follows: (A) *Dx5* vs. *Bx7*; (B) *Ax2** vs. *Bx7*; and (C) *Ax2** vs. *Dx5*. Sequence match criteria was 70% over a 30-bp window. Arrows above main diagonals indicate start codons. Arrows below the main diagonals indicate stop codons. Major gaps in the diagonals are labeled a, b, and c. The short diagonal crossing the main diagonal in C is the region of the inverted repeat insert in the *Ax2** and *Dx5* fragments when compared with the *Bx7* fragment (gap c) and is also labeled c.

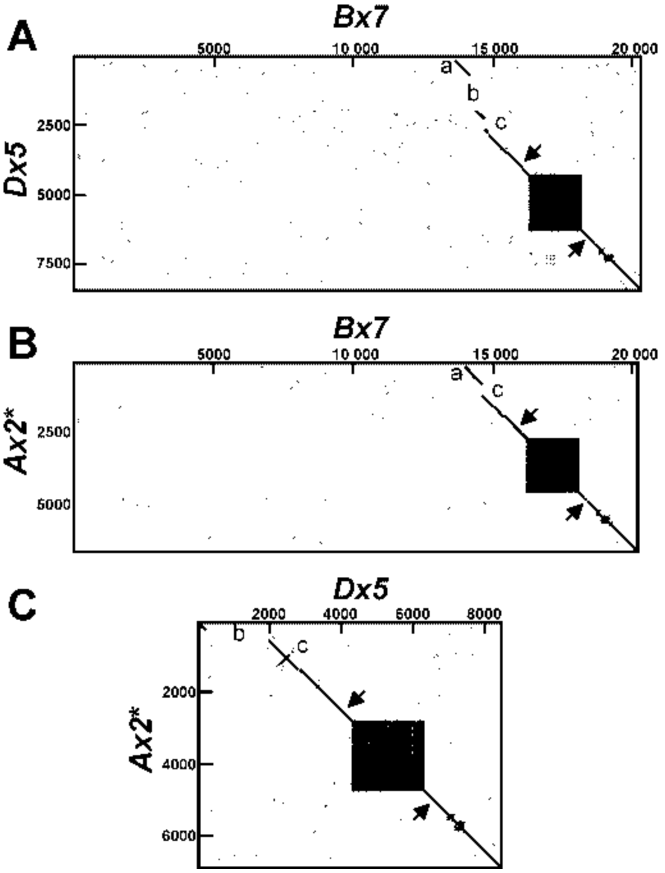
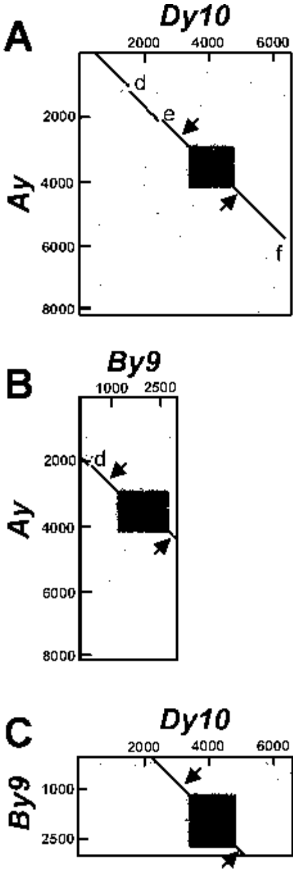


Fig. 2. Dot plot analysis of y-type orthologous HMW-glutenin fragments. The sequence of all three ‘Cheyenne’ y-type orthologous HMW-glutenin gene fragments were compared in all possible combinations as follows: (A) *Ay* vs. *Dy10*; (B) *Ay* vs. *By9*; and (C) *By9* vs. *Dy10*. Sequence match criteria was 70% over a 30-bp window. Arrows above main diagonals indicate start codons. Arrows below the main diagonals indicate stop codons. Gaps in the main diagonal are labeled d, e, and f.



Comparisons between orthologous fragments

Pairwise comparisons among the three *Eco*RI genomic fragments containing the orthologous x-type HMW-glutenin genes are shown in Fig. 1. The block structure within each plot is where the repetitive motifs of the HMW-glutenin repeat domains find multiple matches. The three fragment sequences are similar to one another throughout most of their lengths with two classes of exceptions: (*i*) small differences caused by single base changes and (*ii*) short deletions or duplications and larger insertions. The latter involve the 5′ sequences, relative to the HMW-glutenin non-coding regions, and sequence analysis indicates different classes of transposable-element insertions. For example, the larger gaps, when comparing *Dx5* with the *Bx7* (Fig. 1A; gap b) and *Ax2** fragments (Fig. 1C; gap b), are the result of a retrotransposon insertion event into the *Dx5* fragment (described below). Another likely insertion occurs in the 5′ end of the available *Ax2** and *Dx5* fragments. Although this region is similar in these two fragments (Fig. 1C), the 5′-most ~200 bp of both the *Ax2** and *Dx5* sequences do not match any part of the *Bx7* sequence (Figs. 1A and 1B).

The short diagonal in the 5′ portion of the *Dx5* and *Ax2** comparison in Fig. 1C suggested a miniature, inverted, transposable element (MITE) within the *Dx5* and *Ax2** fragments, but not the *Bx7* fragment. This was confirmed (details below), and explains gap c when the *Bx7* fragment is compared with the *Dx5* and *Ax2** fragments (Figs. 1A and 1B).

The homology matrix comparison of the y-type HMW-glutenin orthologous regions is shown in Fig. 2 (*Ay* vs. *Dy10*, *Ay* vs. *By9*, and *By9* vs. *Dy10*). These fragments are similar, although the *By9* sequence was not extended in the current report. Within the 5′ sequence comparisons there are two short discontinuities. One is due to a 91-bp insertion in the *Ay* sequence (Fig. 2A; gap d) identified as a MITE (details below). The second discontinuity is a 75-bp deletion in the *Ay* sequence, previously reported by Forde et al. (1985), that causes a small gap 5′ to the start codons in the *Ay*–*Dy10* and *Ay*–*By9* comparisons (Figs. 2A and 2B; gap e). Finally, the *Ay*–*Dy10* fragment comparison shows that the final 119 bp of the 3′ end of the *Dy10* fragment has no significant similarity to any portion of the *Ay* fragment (Fig. 2A; gap f). This difference likely indicates another insertion–deletion (indel), but the origin of the difference cannot be determined without at least one additional orthologous sequence in this region.

Paralog divergence

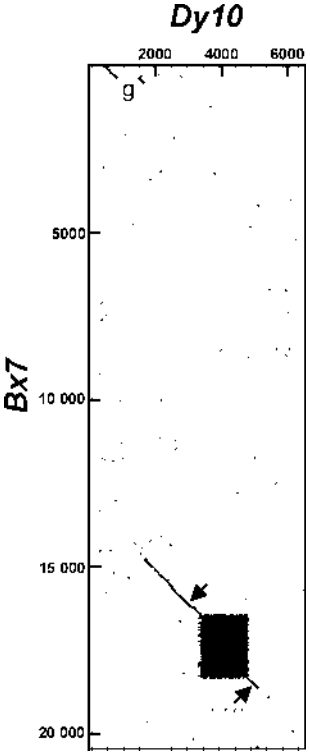
Conservation and divergence in the paralogous x- and y-type HMW-glutenin gene fragments were examined using *Bx7* and *Dy10* as the x- and y-type paralogs, repectively; these two sequences include the most 5′ sequence among the x- and y-type paralogs, respectively. We have previously reported the apparent conservation among the genome-specific paralogous sequences included the conserved sequence that runs in the 5′ direction out to approximately –1200 bp from the start codons through the coding regions, and in the 3′ direction to approximately 200–400 bp from the stop codons, in the vicinity of the polyadenylation sites (Anderson et al. 1998). The extended fragments showed the same conservation pattern near the HMW-glutenin coding regions (Fig. 3), but also indicated conservation of the 5′ *Eco*RI site and the next 380 bp when comparing the *Bx7* and *Dy10* fragments. Two gaps are present from this point up to where the two sequences begin their main alignment, illustrated in Fig. 3. The first gap indicates an additional 900 bp in the *Dy10* sequence (gap g), followed by 14 000 additional bp for the second gap. This implies at least two events occurred in the divergence of the known 5′ paralogous sequences.

Matches of HMW-glutenin fragments to DNA sequence databases

The genomic fragments containing the HMW-glutenin genes were compared with all known sequences deposited at GenBank. In all cases the fragments include the complete HMW glutenin coding DNA (region 4 in Fig. 4), conserved flanking DNA approximately to –1200 bp 5′ and –400 bp 3′ (Anderson et al. 1998), and the extended 5′ and 3′ sequences. The Bx7 clone sequence also includes 16 034 bp 5′ to the start codon and gives, to date, the longest flanking DNA sequences of a wheat gene. As the longest sequences within their respective orthologous sets, the *Bx7* and *Ay* fragments serve as models for those sets. Matches to those two sequences are listed in Table 2 along with the *Dx5* fragment because this fragment contains differences in the 5′ region. The matches are keyed to the fragments diagramed in Fig. 4, with the boxed regions being the longer sequences with significant matches to known plant DNA and wheat ESTs, and with arrowheads indicating shorter matches to ESTs of various cereals. Either the BLASTN or BLASTX *E* values are reported, depending on which indicates the most significant match.

An examination of the 20 425 bp sequence containing HMW-glutenin *Bx7* shows that approximately half of the sequence matches to plant genes and transposable elements. From bp 1288 to 2503 of the *Bx7* fragment (region 1 in Fig. 4) is a BLASTX match to a hypothetical rice (*Oryza sativa* L.) protein (BAB07953; 3e^{-51}) of unknown function. From bp 3098 to 3675 (region 2 in Fig. 4) is another BLASTX match, this time to an *Arabidopsis* putative, non-LTR, retroelement reverse transcriptase from rice (BAA96754; 1e^{-21}).

Fig. 3. Dot plot analysis of paralogous HMW-glutenin gene fragments. Complete sequences of the genomic fragments containing the *Bx7* and *Dy10* paralogous genes were compared using a criterion of a 70% match over a 30-bp window. The arrows above the main diagonal indicate the start codon. The arrows below the main diagonal indicate the stop codon. The smaller gap in the main diagonal is labeled g. The larger gap is unlabeled.

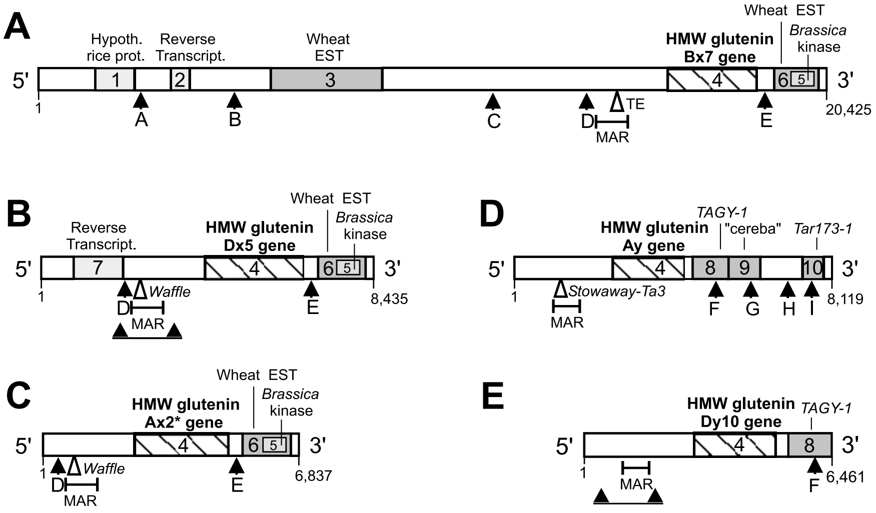


Between regions 1 and 2 is a match (region A) to a wheat EST (BF429214; 3e^{-20}) that itself matches to a barley *cereba* retroelement and may be associated with the retroelement of region 2.

Region B is a 36-bp match to a barley EST (BF258243; 4e^{-7}) that encodes a putative sugar transporter. This region may also be associated with region 3 (Fig. 4) from bp 5009 to 7791, which has a highly significant match to a wheat leaf EST (BE418621; e^{-101}), but has no match to any known gene. The dot blot of BE418621 versus the *Bx7* sequence suggests the exon–intron structure of a gene (Fig. 5A) and analysis of the sequences at the match boundaries suggests exon–intron splice signals (not shown). There are enough base differences between the EST and genomic sequences to indicate that this EST does not originate from the Bx region. It has not yet been determined whether or not this EST belongs to one of the Ax or Dx orthologs or if it is from a closely related gene elsewhere in the wheat genome.

At position 11 571 of the Bx sequence is a 33-bp sequence (c in Fig. 4) matching exactly to DNA within an intron of the barley *Knox3* gene (X83518; 2e^{-7}) and several Triticeae ESTs such as wheat EST BE427704 (2e^{-7}). Closer examination suggests region C is a *Stowaway*-type transposable element with an incomplete TIR (not shown). Similarly, at 14 598 bp is a 33-bp sequence (D in Fig. 4) with a poor BLASTN match (2.4) to a wheat transcription factor (D38111). Analysis of the same region with FASTA shows a

Fig. 4. Cloned *Eco*RI fragments of wheat genomic DNA containing HMW-glutenin genes. HMW-glutenin genes from 'Cheyenne' were cloned as *Eco*RI fragments and the complete DNA sequences of five HMW-glutenin genes were determined. Coding regions for the HMW glutenins are indicated by striped boxes. The fragments containing the three x-type orthologs are shown on top (A, *Bx7*) and at the left (B and C, *Dx5* and *Ax2**, respectively). Fragments containing two of the y-type orthologs are shown on the middle and lower right (D and E, *Ay* and *Dy10*, respectively). Longer matches of flanking DNA to database sequences is shown by light grey boxes for BLASTX matches, and dark grey boxes for BLASTN matches. Shorter DNA sequences having matches are indicated below the fragment by arrowheads and letters (described in Table 2). Small insertion MITE and TE sequences are indicated by open arrowheads. Predicted MARs (Fig. 7) are shown below the sequence by bracketed lines. MAR elements identified in biochemical binding assays are indicated by pointed bars below the MAR label. The initial and final base pair positions are indicated at the ends of the fragments.



longer sequence of 264 bp with a match to the same transcription factor sequence at a much more significant probability ($5e^{-25}$), at 1157 bp 5′ to a barley glucan endo-1,3-β-glucosidase isoenzyme I gene (AF055328; $11e^{-16}$), and a sequence 800 bp 3′ to a barley disease resistance homolog gene (AF166121; $3e^{-18}$). As with region C, region D is similar to known *Stowaway* elements (not shown). More detailed analysis of the wheat *Stowaway* sequences will be reported elsewhere.

Immediately 3′ to *Bx7* is a short match to a sorghum EST (E; BG412266; $3e^{-11}$) of unknown function, followed by an apparent kinase gene. Region 5 has a BLASTX probability score of $2e^{-33}$ to a *Brassica* putative serine (threonine) protein kinase (AAG16628) and only slightly higher probabilities to numerous other kinase-like sequences from *Arabidopsis* and other plants (not shown). Region 5 also has a BLASTN match to a wheat EST (BE499348; $5e^{-7}$). The clone from which this EST was derived was sequenced completely and is closely related to three additional wheat ESTs (mismatch probabilities of 0 to e^{-179}; not shown), several other cereal ESTs (e.g., maize X57273, e^{-58}), and BLASTX matches to numerous *Arabidopsis* kinase sequences with mismatch probabilities as low as e^{-112}. It is not known if this is actually part of an active gene or a remnant of an inactive gene. Further evidence that this region is at least closely related to an active gene is seen by the suggested exon–intron pattern obtained by dot matrix analysis of the matching region of the *Bx7* fragment and full length EST BE499348 (Fig. 5B). As with the 5′-matching EST in Fig. 5A, there is a staggered pattern as expected from comparing an intron-containing gene with a related cDNA without the intron sequences.

The *Dx5* HMW-glutenin *Eco*RI fragment is 8425 bp and mainly contains sequences homologous to the *Bx7* fragment; i.e., the 5′ region of the *Dx5* fragment contains the *Stowaway*-like sequence similar to one found in a wheat transcription factor (region D in Fig. 4), and is found a match in the 5′ end to a sorghum EST encoding a protein of unknown function (region E in Fig. 4) and DNA similar to the derived amino acid sequence of a serine (threonine) protein kinase (region 5 in Fig. 4). Exceptions to the similarity are two significant insertions not found in the *Bx7* fragment. First, unique to the *Dx5* fragment is a 1734-bp insertion (compared with the *Ax* and *Bx* sequences) from bp 256 to1990 of the *Dx5* fragment (region 7 in Fig. 4; gap b in Fig. 1). In this same region the *Dx5* fragment is missing 250 bp common to the *Ax2** and *Bx7* fragments. We speculate the insertion event(s) that distinguished the *Dx5* fragment in this region also deleted these 250 bp. Comparisons of this insertion sequence with known nucleic acid sequences found no significant matches. However, the BLASTX search found matches as significant as $3e^{-71}$ (BAA95894) to non-LTR retroelement reverse transcriptase sequences from *Arabidopsis*. In comparing this sequence with the orthologous regions, the *Dx* insert has a 20–23 bp imperfect TIR at the insertion site that indicates the footprint of a transposon insertion.

Second, at position bp 2265 of the *Dx5* fragment, as with the *Ax2** fragment, 377-bp inverted repeat (insertion c in Fig. 1C) is found when compared with the *Bx7* fragment sequence (gap c in Fig. 1A). Examination of this insertion sequence suggests a new class of plant MITEs,designated *Waffle*. Most of the sequence is composed of a secondary inverted repeat (SIR) allowing potential secondary structure (Fig. 6A). *Waffle* includes no obvious short terminal inverted

Table 2. Sequence database entries related to HMW-glutenin clones.

HMW	Fig. 4[a]	Species[b]	Homology and (or) identity[c]	GenBank accession	BLAST score[d]	Probability[e]
Bx7	1	rice	Hypothetical protein	BAB07953	187 X	$3e^{-51}$
	2	*Arabidopsis*	Reverse transcriptase	BAA96754	92 X	$1e^{-21}$
	3	wheat	EST, unknown	BE418621	379 N	$1e^{-102}$
	4	wheat	*Bx7* HMW-glutenin gene	X13927		
	5	*Brassica*	Serine (threonine) protein kinase	AAG16628	146 X	$2e^{-33}$
	6	wheat	EST, serine (threonine) protein kinase	BE499348	62 N	$5e^{-7}$
	A	wheat	EST, *cereba* retroelement	BF429214	155 N	$3e^{-20}$
	B	barley	EST, putative sugar transport	BF258243	64 N	$4e^{-7}$
	C	barley	*Stowaway*-like in *Knox3* gene	X83518	66 N	$2e^{-7}$
	D	wheat	*Stowaway*-like in *TF-HBP-1a*[f]	D38111	42 N	2.4
	E	sorghum	EST, unknown	BG412266	76 N	$3e^{-11}$
Dx5	7	*Arabidopsis*	putative reverse transcriptase	BAA95894	182 X	$3e^{-71}$
	4	wheat	*Dx5* HMW-glutenin gene	X12928		
	5	*Brassica*	Serine (threonine) protein kinase	AAG16628	145 X	$5e^{-33}$
	6	wheat	EST, serine (threonine) proten kinase	BE499348	62 N	$5e^{-7}$
	D	wheat	*Stowaway*-like in *TF-HBP-1a*[f]	D38111	64 N	$6e^{-7}$
	E	sorghum	EST, unknown	BG412266	76 N	$1e^{-10}$
Ay	4	wheat	*Ax2** HMW-glutenin gene	X03042		
	8	barley	*BAGY-2* LTR retroelement	AF254799	288 N	$1e^{-74}$
	9	barley	*cereba*-like retroelement	AF078801	353 N	$4e^{-94}$
	10	rye	*R173-1* dispersed repeat	X64100	208 N	$2e^{-50}$
	F	barley	EST, *BAGY-2* retroelement	BE422035	62 N	$9e^{-7}$
	G	wheat	EST, *cereba* retroelement	BE398545	450 N	$1e^{-124}$
	H	barley	EST, *BAGY-2* retroelement	BG418379	389 N	$1e^{-105}$
	I	barley	EST, *R173-1* dispersed repeat	BG310371	316 N	$3e^{-83}$

Note: The Ay and Bx7 clones are the longest sequences among their orthologs and are shown first. Extended sequence matches are listed by number, followed by short or lower probability matches denoted by letters. The highest scores are self-matches in GenBank. Homology and (or) identity is determined as the highest match via either BLASTN or BLASTX. X, BLASTX; N, BLASTN.

[a] Keyed to Fig. 4.

[b] Common name for cereal species.

[c] Highest similarity match with BLAST.

[d] Score for highest match.

[e] Probability match is by chance.

[f] HBP-1a transcription factor gene.

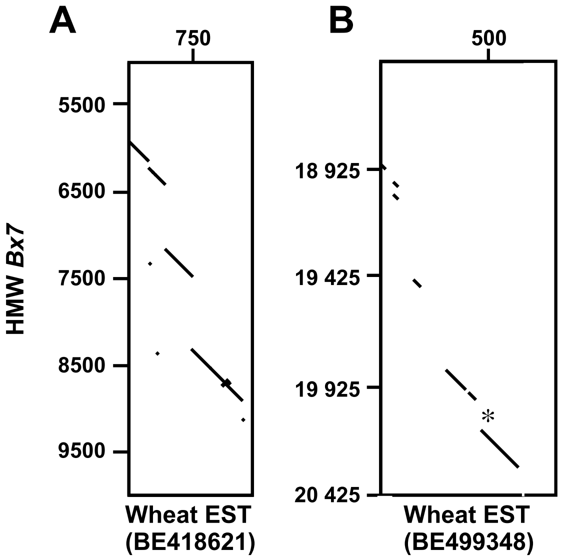
repeat (TIR) characteristics of the plant MITE classes such as the maize (*Zea mays* L.) *Tourist* (Bureau and Wessler 1992) and *Stowaway* (Bureau and Wessler 1994), and therefore is similar to rice MITE classes *Pop* and *Crackle* (Song et al. 1998). *Waffle* also includes an 18-bp imperfect insertion site duplication, compared with the 8-bp duplication of the insertion site for *Pop* and *Crackle* or 2-bp duplication of *Tourist* and *Stowaway*.

The *Ax2** *Eco*RI fragment is 6837 bp long and is closely homologous throughout its sequence to orthologous *Bx7* and *Dx5* fragments with the exception of region 7, which is unique to the *Dx5* fragment.

Measuring 8119 bp, the Ay fragment is the longest available sequence among the y-type HMW-glutenin genes, and is different from the x-type sequences by having mainly non-homologous sequences 5′ to the promoter region and a retrotransposon-associated region 3′ to the HMW-glutenin coding region (Figs. 1 and 4). At position 1093 of the *Ay* fragment is a 91-bp sequence with MITE characteristics: duplication of the two base pair (TA) insertion site, short TIRs, and an SIR with the potential to form secondary structures (Fig. 6B). The TIR is almost identical to the consensus TIR of the maize MITE family *Stowaway* (CTCCCTCCGTT), and this wheat MITE has been named *Stowaway*-Ta3, consistent with two wheat *Stowaway* elements previously described (Bureau and Wessler 1994).

The *Ay* terminal 4000 bp of the 3′ end contains sequences that match Triticeae repetitive and transposable elements. These include a sequence (region 8) from approximately bp 4500 to 5600, whose best match is to the *BAGY-2* retrotransposon LTR from barley (AF254799; e^{-74}; Shirasu et al. 2000) and represents a new class of retrotransposons related to the *BAGY* (barley *gypsy*-like) class of barley. To remain consistent with the barley elements, this new wheat class is named *TAGY-1* (first member of *Triticum aestivum gypsy*-like retrotransposons). Region 9, from bp 5700 to 6500, matches nearby sequences of both *cereba*, a Ty3-*gypsy* related element reported in barley ($4e^{-94}$; Presting et al. 1998) to be localized to the centromere, and *Sabrina*, a retroelement reported from the vicinity of the barley *Rar1* locus ($3e^{-74}$; Shirasu et al. 2000). However, these two matches are not within the previously reported identified elements and have low similarity to the R173-3 rye dispersed-repeat sequence (Rogowsky et al. 1992). It is not clear if region 9 represents a new class of wheat retrotransposons, and it is referred to here as *cereba*-associated, awaiting further characterization. Finally, the sequence from approximately bp 7300 to 7800 (region 10) of the 3′ sequence

Fig. 5. Dot matrix analysis of HMW-glutenin *Bx7* fragment and wheat ESTs. Regions of the *Bx7* fragment DNA sequence with significant matches to wheat ESTs were analyzed by dot matrices to the full-length sequences of those respective ESTs. A match criterion of 70% over a 30-bp window was used. (A) 5000 to 10 000 bp of the *Bx7* fragment vs. EST BE418621. (B) 18425 to 20425 bp of the *Bx7* fragment vs. EST BE499348. The asterisk indicates a gap coinciding with an intron in a *Brassica* serine (threonine) protein kinase gene (AAG16628).



of the *Ay* fragment matches the dispersed, retrotransposon-like, repeated element R173-1 LTR region, previously believed specific to rye ($2e^{-50}$; Rogowsky et al. 1992). This third new class of wheat elements is named *Tar173-1* (first reported member of a family of *Triticum aestivum* retroelements similar to the rye *R173-1* sequence).

All three transposon – repetitive regions in the 3′ flank of the *Ay* gene match to Triticeae ESTs or can be indirectly associated with Triticeae ESTs. *TAGY-1* has a low significance match ($2e^{-6}$) to the barley endosperm EST BE422035 (F in Fig. 4). This EST is related by BLASTX to a putative *Arabidopsis* retroelement polyprotein sequence (BAB02143; $5e^{-26}$), by BLASTN to *BAGY-1* (AF254799; 0.0), and to other Triticeae ESTs with matches with probabilities as significant as $7e^{-32}$ (BE414408). The *cereba*-like region has a highly significant match to a wheat EST BE398545 (G in Fig. 4; e^{-124}) plus four other wheat and barley ESTs at $3e^{-19}$ to $7e^{-5}$ (not shown). Wheat EST BE398545, in turn, matches best to *cereba* and *Sabrina*. The *Tar173-1* sequence is most similar to a barley EST (BG310371; $3e^{-83}$). Finally, region H matches to a barley EST (BG418379; e^{-105}) that itself matches best with a barley *BAGY-2* retroelement.

The complete *Dy10 Eco*RI fragment sequence (6461 bp) is shorter than the *Ay* fragment, but with an additional 480 bp at its 5′ end when compared with the *Ay* fragment, and about 2200 bp shorter at the 3′ end than the *Ay* fragment. The *Dy10* fragment is similar within the orthologous sequences available, with the exception of the *Stowaway*-Ta3 element found in the *Ay* fragment, and the final 119 bp of the *Dy10* fragment does not match the *Ay* fragment. The latter is likely due to another insertion that is only partially represented in either the *Ay* or *Dy10* fragment. The *By* orthologous sequence could be used to determine the origin of the difference.

Matrix attachment regions in gene flanking DNA sequences

Matrix attachment regions (MARs) are DNA sequence elements that associate with the protein scaffold of the nuclear matrix and are believed to be related to gene activity by involvement in creating chromatin domains competent for transcription (Jenuwein et al. 1997; Allen et al. 2000). To examine the presence of such elements within the sequenced regions around the wheat HMW-glutenin genes, the predictive program MarFinder was applied to the five extended DNA sequences of the wheat HMW-glutenin genes (Fig. 7). All five HMW-glutenin fragments include a predicted MAR in the 5′ flanking DNA, approximately centered 1300 bp upstream of the start codon. To test these predictions, a biochemical test of DNA binding to nuclear matrix preparations was carried out for *Dx5* and *Dy10* (Fig. 8). Restricted HMW-glutenin clones were incubated with wheat nuclear matrix preparations. The mixture was centrifuged to determine what DNA fragments remain unbound in the soluble fraction and which DNA fragments bind to the nuclear matrices and enter the pellet. Both HMW-glutenin fragments include a restriction fragment that preferentially binds the nuclear matrix preparations (Fig. 8, arrowheads), and in both cases these restriction fragments are coincident with the regions of predicted MAR activity (Fig. 7, black boxes; Fig. 4, bracketed and pointed lines 5′ to all HMW-glutenin coding regions). In addition, Fig. 7C shows the location of a biochemically determined weak MAR binding for the *Bx* gene from wheat ‘Glenlea’ (Rampitsch et al. 2000).

The detection of MARs near the wheat HMW-glutenin gene raises the question of whether or not such MARs may exist adjacent to other classes of wheat storage protein genes, such as gliadins. We have also previously sequenced extended regions of several members of other classes of wheat storage protein genes including α-gliadins (Anderson et al. 1997), γ-gliadins (Anderson et al. 2001), and ω-gliadins (Hsia and Anderson 2001). The longest available sequences for these gene families all show predicted MARs within both the immediate 5′ and 3′ flanking DNAs (Fig. 9), although the predictions are not as significant as for the HMW-glutenin genes.

Discussion

The wheat HMW-glutenin subunits are critical for wheat quality. These polypeptides and their encoding genes are, respectively, the most studied from any of the cereal crops. To understand more aspects of the control of these genes and the structure of the wheat genome, we have extended the known DNA sequence flanking five orthologous and paralogous members of the wheat HMW-glutenin gene family. Within these extended sequences is evidence of previously unreported, closely linked, non-glutenin genes, a number of retrotransposons and MITE insertions that form the main mechanism for the early stages of divergence of these sequences, and predicted and detected MAR elements.

Although the wheat and barley genomes are, respectively, 40 and 13 times the size of the rice genome, linkage data indicates that wheat and barley have gene-rich islands

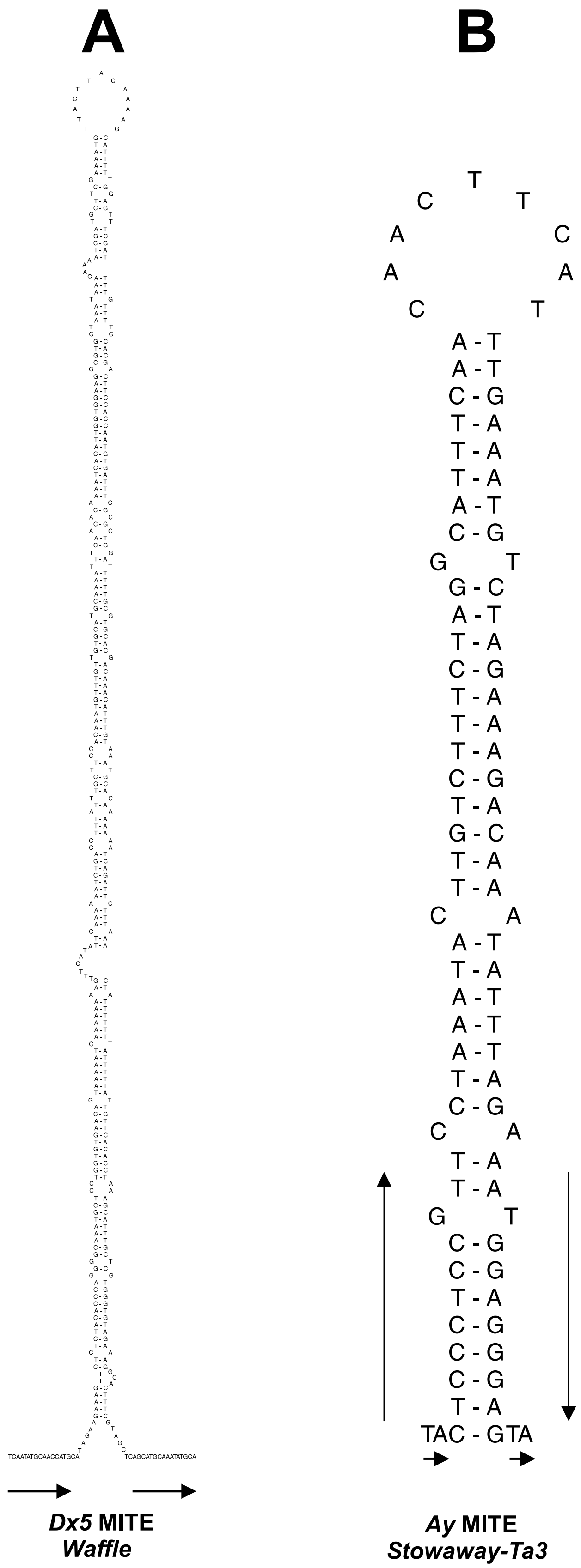


Fig. 6. MITEs in HMW-glutenin gene fragments. DNA sequence fragments from HMW-glutenin gene clones suggesting MITEs are shown in potential secondary structure alignment. (A) 377-bp MITE (*Waffle*) from the *Dx5* fragment. (B) 91-bp MITE (*Stowaway*-Ta3) from the *Ay* fragment.

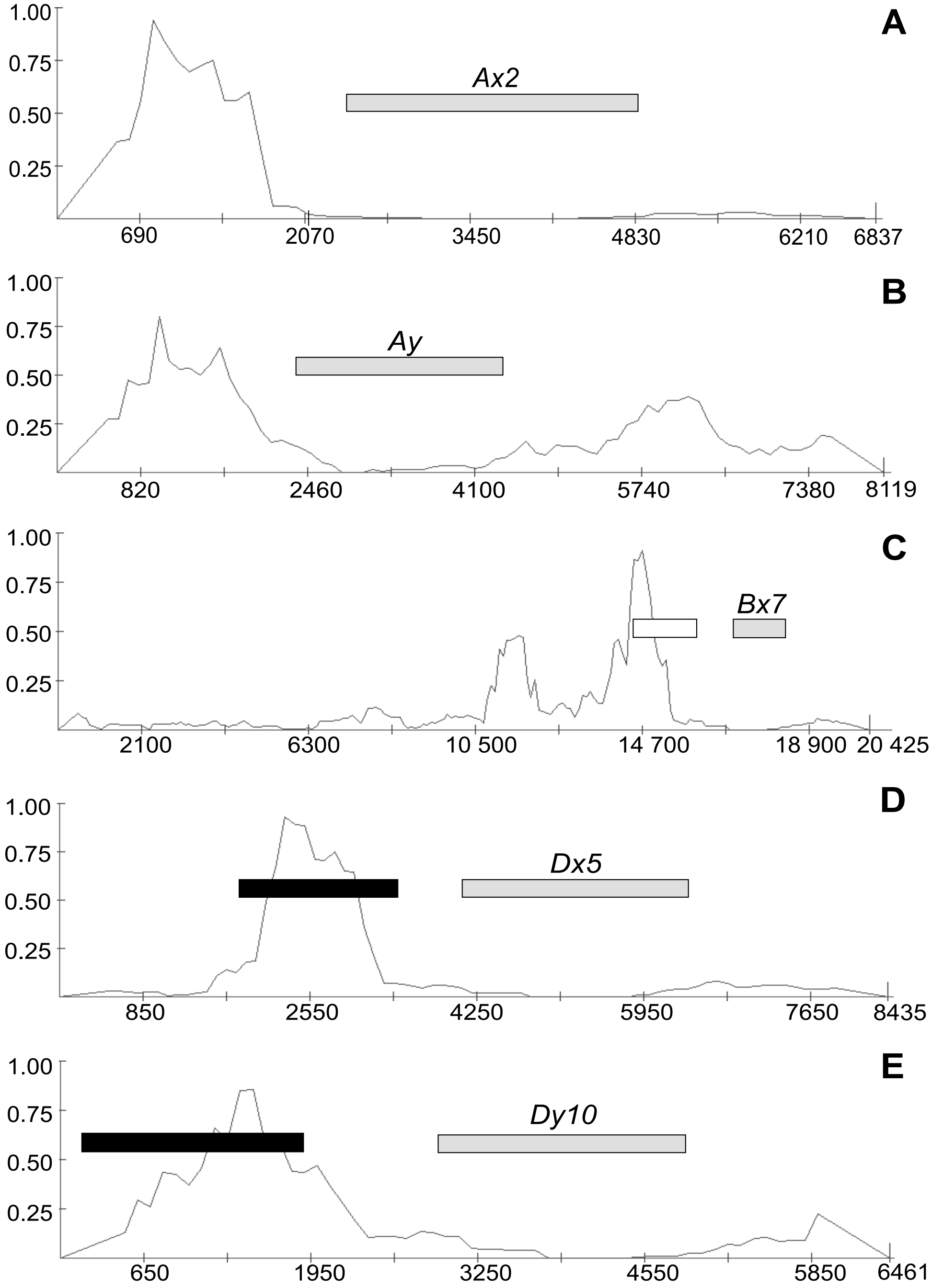
interspersed with long stretches of DNA with few or no genes (Feuillet and Keller 1999; Devos and Gale 2000). The spacing of genes was suggested to be about one gene per 5–15 kb, significantly less than the expected 100–250 kb if Triticeae genes are randomly and evenly spaced. Similarly, available data from maize and rice indicates gene spacings of 5–25 kb and 5–10 kb, respectively (Feuillet and Keller 1999).

The longest contiguous Triticeae genomic sequences reported are barley genomic intervals of 60 kb (Panstruga et al. 1998) and 66 kb (Shirasu et al. 2000). In both examples, three genes were identified within each of these intervals, but were relatively clustered within 18 and 22 kb, respectively. Large portions of the remainder of the reported barley sequences were occupied by retrotransposons. Although the contiguous wheat sequences in the current report are shorter, the same pattern is suggested: genes relatively clustered and clusters interspersed with transposon-rich regions. The four genes in the *Bx7* fragment are a hypothetical rice gene, a gene of unknown function immediately 5′ to *Bx7*, *Bx7* itself, and a putative serine (threonine) protein kinase gene immediately 3′ to the HMW-glutenin gene (Fig. 4). It is not known if the three non-glutenin genes are themselves active genes, but they are at least remnants of active genes because, in at least two cases, ESTs are related to these genomic sequences and a comparison of both of these ESTs to the *Bx7* genomic fragment indicates the exon–intron structure of genes (Fig. 5). In addition, the suggested exon–intron structure of the *Bx7* and EST BE418621 comparison contains six of eight potential splice junctions (Fig. 5A) containing the canonical GT–AG elements at the putative intron termini (not shown). Although the wheat EST BE499348 does not match as closely to the *Bx7* fragment as EST BE418621, one of the putative introns (Fig. 5B, asterisk) coincides exactly with an intron in a Brassica serine (threonine) protein kinase gene (AAG16628).

Although the *Bx7* fragment sequence supports the concept of Triticeae gene islands, the extent of such islands and the size and distribution of retrotransposon blocks such as have been found in maize is still to be determined. These are critical questions in any future effort to sequence large sections of the Triticeae and maize genomes. It is commonly stated that sequencing such large plant genomes is beyond practicality considering current resources and technology, but this may not be true if most of the genes are clustered and strategies can be developed to concentrate sequencing only in such clusters. Similarly, wheat's multiple genomes may not be as serious an obstacle as sometimes predicted if the genomes are similar enough to allow focusing on diploid members of the Triticeae tribe such as barley or the diploid ancestors of wheat.

A study of orthologous regions such as surrounding the HMW-glutenin genes also have the potential to address issues related to plant genome evolution. An understanding of the

Fig. 7. Predicted and detected MARs in HMW-glutenin genes. Complete extended DNA sequences of five HMW-glutenin genes were analyzed for predicted MAR elements. Using the MarFinder program at NCGR, MAR potentials are normalized to 1.0 and plotted versus the center of the sequence window. Grey bars indicate coding regions for HMW-glutenin genes. Black bars in the *Dx5* and *Dy10* HMW-glutenin frames indicates restriction fragments testing positive for MAR elements in Fig. 8. (A) *Ax2** fragment. (B) *Ay* fragment. (C) *Bx7* fragment. (D) *Dx5* fragment. (E) *Dy10* fragment. The open bar in C is the MAR sequence identified by Rampitsch et al. (2000).



precise early and intermediate steps in the fine structure details of genome evolution is unlikely to be resolved by comparison among species whose intergenic sequences have already completely diverged (Avramova et al. 1996; Chen et al. 1996). We are interested in studying orthologous and paralogous regions of Triticeae DNA sequence to gain evidence of the early events leading to divergent genome structure. The three genomes of hexaploid wheat offer a convenient system when combined with the more distantly related members of the Triticeae grass tribe such as barley and rye. The HMW glutenins of wheat are a useful system because there are two paralogous genes in each genome (the x- and y-type genes), and three homoeologous genomes in the hexaploid bread wheats. Comparison of the x- and y-type fragments indicates the results of events after the duplication that created these paralogous types, and results indicate that the major mechanism of divergence is the insertion of retrotransposons.

Within each of the two HMW-glutenin paralogous gene sets there is conservation over most of the known orthologous DNA sequences except for transposon insertions. Within the x-type orthologs, one insertion is shared by the *Dx* and *Ax* fragments (Fig. 4, MITE *Waffle*), whereas a second insertion event is unique to the *Dx* fragment (Fig. 4, region 7). The former suggests that the A and D genomes shared a common ancestor after splitting from the B genome lineage, and that a later insertion occurred in the *Dx* fragment.

Similarly, the 5′ terminal sequence of *Bx7* does not match the sequence shared by the *Ax2** and *Dx5* sequences and is likely the result of an uncharacterized transposition event into either the *Bx7* ancestor or the shared ancestor of the *Ax* and *Dx* fragments. More distal 5′ sequences are necessary to resolve the two possibilities.

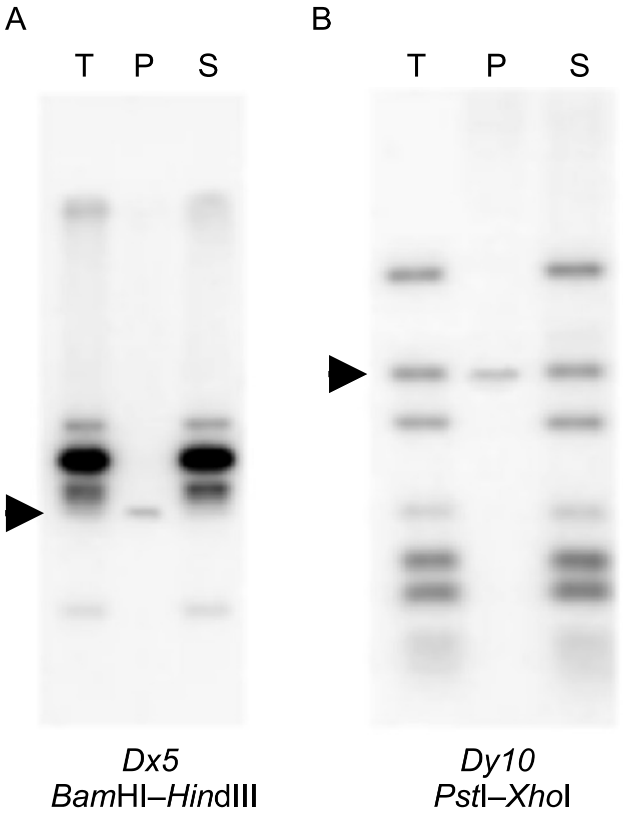
Although less data is available, the y-types are also similar to each other over their known sequence. One exception is the insertion of a MITE (*Stowaway*-Ta3) in the 5′ region of the *Ay* fragment. A second exception is that the final 119 bp of *Dy10* does not align with the appropriate *Ay* sequence, indicateing another indel event after the *Ay* and *Dy* orthologs diverged.

Plant genes are commonly associated with transposable elements; LTR and non-LTR retrotransposons such as LINES and SINES (Kumar and Bennetzen 1999), and MITEs (Wessler 1998). In maize, regions of nested retrotransposons comprise a major portion of the chromosomal DNA sequence thus far reported (SanMiguel et al. 1996). The largest reported transposon associated with wheat-specific genomic DNA is the *WIS2-1A copia*-like 8.6-kb insertion into a HMW-glutenin gene (Murphy et al. 1992). Other reports of wheat transposable elements include transposon-like sequences in a wheat glutathione *S*-transferase gene (Mauch et al. 1991), and a small transposon-related sequence in a gene promoter (Anderson et al. 1998). Surveys of small genomic fragments, mainly from PCR studies, have reported that wheat contains LTR retrotransposons of the *gypsy* (Kumekawa et al. 1999) and *copia* types (Hirochika and Hirochika 1998; Gribbon et al. 1999; Matsuoka and Tsunewaki 1999*a*), as well as non-LTR retrotransposons (Matsuoka and Tsunewaki 1999*b*; Noma et al. 1999).

The flanking sequences to the HMW-glutenin genes show both evidence of copies of retrotransposons and MITEs related to previously reported cereal transposons and also likely novel elements. The 5′ flanks of the x-type genes contain retroelement-related sequences although no evidence of LTRs has been found in the *Bx7* and *Dx5* fragments, indicating regions 2 and 7 may be associated with one or more non-LTR retrotransposons.

The close proximity of the two *Ay*-fragment retrotransposon-related and *Tar173-1* sequences may be due to nested insertions such as has been seen in maize (San Miguel et al. 1996). The *Tar173-1* sequence is closely related to the rye R173 dispersed repeat family that has a more distant relationship to several Triticeae retrotransposons. The exact relationship is unclear, but *Tar173-1* is likely one of the more abundant repeat elements in wheat because, although there is relatively little genomic sequence available for wheat, another sequence related to R173-1 is found just 3′ to the wheat α-gliadin gene CNN18 (U51305; Anderson et al. 1997). Because all of the retrotransposon-related elements in the 3′ sequence flanking *Ay* have matching wheat and barley ESTs, these families of retroelement-type sequences are active in transcript production in the Triticeae.

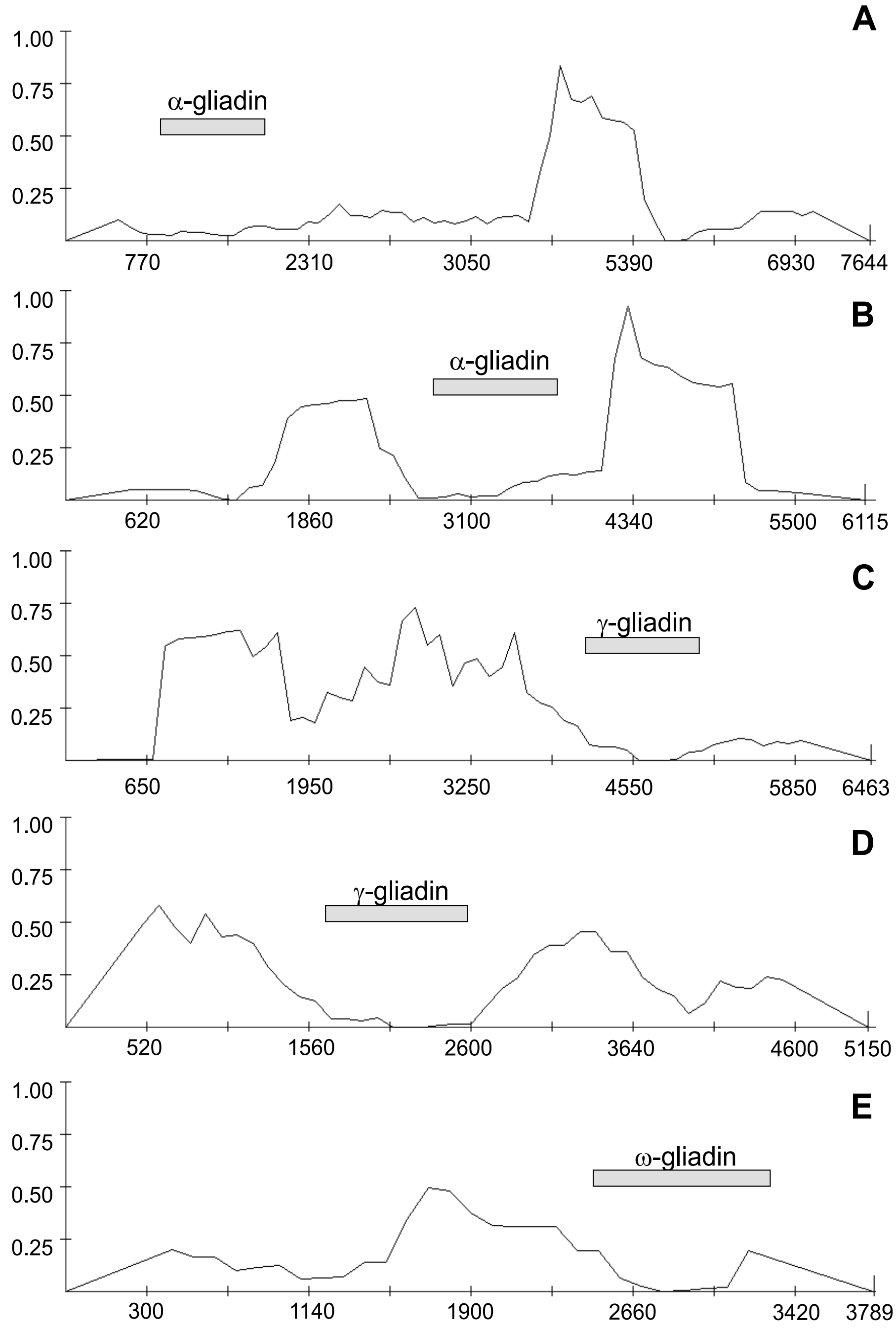
Fig. 8. MARs of the D genome HMW-glutenin genes. The two fragments containing the Dgenome HMW-glutenin genes *Dx5* (A) and *Dy10* (B) were restricted with the indicated restriction enzymes and the digestion products were tested for binding to wheat nuclear matrix preparations. T, total original fragments; P, pellet fraction; S, soluble fraction. Arrows indicate the single restriction fragment from the original clones that bound to the matrix preparation: *Dx5*, 1659-bp *Bam*HI–*Hin*dIII; *Dy10*, 1764-bp *Pst*I–*Xho*I. Gels A and B were electrophoresed for different lengths of time.



Transposon activity alone may not account for all divergences among the HMW-glutenin orthologs and paralogs. The approximately 14 000-bp discontinuity in the conservation of the *Bx7* and *Dy10* fragment 5′ sequences includes two gene-like sequences (Fig. 4, regions 1 and 3). Two possibilities are that (*i*) the entire region of this discontinuity was translocated into the region 5′ to an ancestral x-type HMW-glutenin gene or (*ii*) a deletion of this region occurred in a y-type ancestor.

Another class of transposable elements found flanking the HMW-glutenin genes are MITEs; insertion sequences with short length (80–500 bp), TIRs of 10–15 bp, SIRs, and the potential for forming secondary structures. The HMW-glutenin *Ay* fragment contains a 91-bp MITE that is a member of the *Stowaway* class (*Stowaway*-Ta3). *Stowaway* elements seem

Fig. 9. Predicted MARs in wheat gliadin genes. MAR element prediction was carried out on the longest available DNA sequence for three classes of wheat gliadin genes: two α-gliadins, two γ-gliadins, and one ω-gliadin. MAR potentials were determined as in Fig. 7. Grey bars indicate coding regions for these gliadin genes. (A) α-Gliadin αCNN10. (B) α-Gliadin αCNN5. (C) γ-Gliadin γ2656. (D) γ-Gliadin G6. (E) ω-Gliadin ωF20b. Grey bars indicate coding regions.



to be common in the Triticeae. A 60-kb barley sequence contained two *Stowaway* class elements (Panstruga et al. 1998) and elements are present in numerous ESTs of wheat, barley, and rye (not shown). Related *Stowaway*-Ta3 elements are found near several Triticeae genes, including 250 bp, 3′ of the stop codon of the wheat AWJL175 gene (X81368); 300 bp, 5′ to the start codon of a wheat metallothionein II gene (X68288); and 700 bp, 5′ to the start codon of a barley starch synthase I gene (AF234163).

The *Ax–Dx* MITE (*Waffle*) differs from both the *Ay* MITE (*Stowaway*-Ta3) and previously reported MITEs. *Waffle* was only found in a single currently known Triticeae EST (BE419296; $2e^{-13}$), originating from a wheat root library. In turn, this EST is as yet unique and has no significant similarity to other ESTs. This suggests that *Waffle* is either a less active MITE or that *Stowaway*-Ta3 favors the vicinity of coding regions more than *Waffle* does. *Waffle* is similar to the rice MITEs *Pop* and *Crackle* (Song et al. 1998) with no immediate TIR, and with one SIR.

MARs associated with the HMW-glutenin genes are apparently hot-spots for wheat MITE insertion. Within a 500-bp portion of the MAR regions of the HMW-glutenin genes are the insertion sites for two MITEs (*Stowaway-Ta3* and *Waffle*), the *Stowaway*-related region D, plus the unclassified small transposition event reported for a *Bx* allele (Anderson et al. 1998) (Fig. 4).

MITEs have been shown to be associated with plant genes, not randomly distributed throughout the plant genome, and occur mainly in promoter regions, 3′-UTRs, and introns (Wessler 1998). It has also been speculated that because MITE DNA sequences often resemble short 5′ and 3′ DNA expression and stability control elements, they might be involved in the evolution of gene function (Wessler 1998).

MARs are located in the 5′ proximal regions of all HMW-glutenin genes. This is supported in the present report by the biochemical analysis of the two D-genome HMW-glutenin genes, the report by Rampitsch et al. (2000) for *Bx*, and the predicted location of MARs in five HMW-glutenin gene extended sequences reported here. In addition, MARs are predicted to be present in the extended gene sequences of wheat α-, γ-, and ω-gliadin genes, although for the gliadin genes the MARs are predicted both 5′ and 3′ to the coding regions. Although the expression status of all the HMW-glutenin genes is known, there is usually no direct information to determine if any specific gliadin gene is active because the gliadin families are so large. In at least one case, the G6 γ-gliadin gene (Fig. 9) is a pseudogene. The close proximity of MARs to the wheat prolamine genes may be related to the expression of these genes at very high levels during grain filling. A similar speculation relates to the lack of introns in the cereal prolamine genes.

Rampitsch et al. (2000) identified a MAR-like sequence upstream of the HMW glutenin from the *Bx* gene of ‘Glenlea’. They localized the element at about –750 bp to –1560 bp 5′ from the start codon Fig. 7). The 820-bp sequence tested by Rampitsch et al. (2000) showed no tissue-specificity in transient expression of bombarded tissues, but may have supported an increase in expressing foci in the bombarded tissue. Because these were transient expression experiments, it is not known what the effect of stable incorporation of HMW-glutenin MARs would be on chromatin organization and associated gene expression. The broad peaks of predicted MAR structure (Fig. 8) may indicate that longer stretches of DNA are needed for full MAR functionality. Blechl et al. (1998) have speculated that the high levels of stable expression of transgenic HMW-glutenin genes may be due to fortuitous inclusion of MAR sequences in the transforming gene constructs.

MAR predictive algorithms attempt to include a number of factors, such as the origin of replication signals, T–G rich regions, curved DNA, kinked DNA, A–T richness, and topoisomerase II sites (Singh et al. 1997). The predicted MAR regions overlap the 5′ divergence of the x- and y-type HMW-glutenin genes and include sequences that do not match between the types. MARs are not known to be determined by specific conserved sequences, but by an as yet unclear summation of DNA elements. Thus, although some of the specific DNA sequence has diverged between MAR regions of the *Dx*- and *Dy*-type HMW-glutenin genes, the MAR functionality has been maintained. The dissection of the HMW-glutenin MAR regions in transgenic wheat assays will better define the limits and activities of these MARs.

References

Allen, G.C., Hall, G.E., Childs, L.C., Weissinger, A.K., Spiker, S., and Thompson, W.F. 1993. Scaffold attachment regions increase reporter gene expression in stably transformed plant cells. Plant Cell, **5**: 603–613.

Allen, G.C., Spiker, S., and Thompson, W.F. 2000. Use of matrix regions (MARs) to minimize transgene silencing. Plant Mol. Biol. **43**: 361–376.

Anderson, O.D., and Greene, F.C. 1989. The characterization and comparative analysis of high M_r glutenin genes from genomes A and B of a hexaploid bread wheat. Theor. Appl. Genet. **77**: 689–700.

Anderson, O.D., Yip, R.E., Halford, N.G., Forde, J., Shewry, P.R., Malpica-Romero, J.-M., and Greene, F.C. 1989. Nucleotide sequences of two high-molecular-weight glutenin subunit genes from the D-genome of a hexaploid bread wheat, *Triticum aestivum* L. cv. Cheyenne. Nucleic Acids Res. **17**: 461–462.

Anderson, O.D., Litts, J.C., and Greene, F.C. 1997. The γ-gliadin gene family: I. Characterization of ten new γ-gliadin genomic clones, evidence for limited sequence conservation of flanking DNA, and southern analysis of the gene family. Theor. Appl. Genet. **95**: 50–58.

Anderson, O.D., Abraham-Pierce, F.A., and Tam, A. 1998. Conservation in wheat high-molecular-weight glutenin promoter sequences: comparisons among loci and among alleles of the *GLU-B1–1* locus. Theor. Appl. Genet. **96**: 568–576.

Anderson, O.D., Hsia, C.C., and Torres, V. 2001. The wheat γ-gliadin genes: characterization of ten new sequences and further understanding of γ-gliadin gene family structure. Theor. Appl. Genet. **103**: 323–330.

Avramova, Z., Tkihonov, A., SanMiguel, P., Jin, Y.-K., Liu, C., Woo, S.S., Wing, R.A., and Bennetzen, J.L. 1996. Gene identification in a complex chromosomal continuum by local genomic cross-referencing. Plant J. **10**: 1163–1168.

Blechl, A.E., and Anderson, O.D. 1996. Expression of a novel high-molecular-weight glutenin subunit in transgenic wheat. Nat. Biotechnol. **14**: 875–879.

Blechl, A.E., Le, H.Q., and Anderson, O.D. 1998. Engineering changes in wheat flour by genetic transformation. J. Plant Physiol. **152**: 703–707.

Bureau, T.E., and Wessler, S.R. 1992. *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell, **4**: 1283–1294.

Bureau, T.E., and Wessler, S.R. 1994. *Stowaway*: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Plant Cell, **6**: 907–916.

Chen, M., SanMiguel, P., De Oliveira, A.C., Woo, S.-S., Zhang, H., Wing, R.A., and Bennetzen, J.L. 1996. Microlinearity in *sh2*-homologous regions of the maize, rice, and sorghum genomes. Proc. Natl. Acad. Sci. U.S.A. **94**: 3431–3435.

Depicker, A., and van Montagu, M. 1997. Post-transcriptional genes silencing in plants. Curr. Opin. Cell. Biol. **9**: 373–382.

Devos, K.M., and Gal,e M.D. 2000. Genome relationships: the grass model in current research. Plant Cell, **12**: 637–646.

Feuillet, C., and Keller, B. 1999. High gene density is conserved at syntenic loci of small and large grass genomes. Proc. Natl. Acad. Sci. U.S.A. **96**: 8265–8270.

Forde, J., Malpica, J.-M., Halford, N.G., Shewry, P.R., Anderson, O.D., and Green, F.C. 1985. The nucleotide sequence of a HMW glutenin subunit gene located on chromosome 1A of wheat (*Triticum aestivum* L). Nucleic Acids Res. **13**: 6817–6832.

Gribbon, B.M., Pearce, S.R., Kalendar, R., Schulman, A.H., Paulin, L., Jack, P., Kumar, A., and Flavell, A.J. 1999. Phylogeny and transpositional activity of *Ty1*-copia group retrotransposons in cereal genomes. Mol. Gen. Genet. **261**: 883–891.

Halford, N.G., Forde, J., Anderson, O.D., Greene, F.C., and Shewry, P.R. 1987. The nucleotide and deduced amino acid sequences of an HMW glutenin subunit gene from chromosome 1B of bread wheat (*Triticum aestivum* L) and comparison with those of genes from chromosomes 1A and 1D. Theor. Appl. Genet. **75**: 117–126.

Hall, G.E. Jr., and Spiker, S. 1994. Isolation and characterization of nuclear scaffolds. *In* Plant molecular biology manual. *Edited by* S.B. Gelvin and R.A. Schilperoort. Kluwer Academic Press, Dordrect, The Netherlands. pp. 1–12.

Hirochika, H., and Hirochika, H. 1998. *Ty1*-copia group retrotransposons as ubiquitous components of plant genomes. Jap. J. Genet. **68**: 35–46.

Hsia, C.C., and Anderson, O.D. 2001. Isolation and characterization of wheat ω-gliadin genes. Theor. Appl. Genet. **103**: 37–44.

Jenuwein, T., Forrester, W.C., Fernandez-Herrero, L.A., Laible, G., Dull, M., and Grosschedl, R. 1997. Extension of chromatin accessibility by nuclear matrix attachment regions. Nature (London), **385**: 269–272.

Kolster, P., Krechting, C.F., and van Gelde, W.M.J. 1993. Expression of individual HMW glutenin subunit genes of wheat (*Triticum aestivum* L) in relation to differences in the number and type of homoeologous subunits and differences in genetic background. Theor. Appl. Genet. **87**: 209–216.

Kramer, J.A., Singh, G.B., and Krawetz, S.A. 1996. Computer assisted search for sites of nuclear matrix attachment. Genomics, **33**: 305–308.

Kumar, A., and Bennetzen, J.L. 1999. Plant retrotransposons. Ann. Rev. Genet. **33**: 479–532.

Kumekawa, N., Ohtsubo, H., Horiuchi, T., and Ohtsubo, E. 1999. Identification and characterization of novel retrotransposons of the gypsy type in wheat. Mol. Gen. Genet. **260**: 593–602.

Macritchie, F. 1992. Physiochemical properties of wheat proteins in relation to functionality. Adv. Food Nutr. Res. **36**: 1–87.

Marchylo, B.A., Lukow, O.M., and Kruger, J.E. 1992. Quantitative variation in high molecular weight glutenin subunit 7 in some Canadian wheats. J. Cereal Sci. **15**: 29–37.

Matsuoka, Y., and Tsunewaki, K. 1999*a*. Wheat retrotransposon families revealed by analysis of reverse transcriptase domain. Mol. Biol. Evol. **13**: 1384–1392.

Matsuoka, Y., and Tsunewaki, K. 1999*b*. Detection and analysis of non-LTR retrotransposons in wheat. Proceedings of the 9th International Wheat Genetics Symposium, 2–7 August 1998, Saskatoon, Sask. *Edited by* A.E. Slinkard. University Extension Press, University of Saskatchewan. pp. 24–27.

Mauch, F., Hertig, C., Rebmann, G., Bull, J., and Dudler, R. 1991. A wheat glutathione *S*-transferase gene with transposon-like sequences in the promoter region. Plant Mol. Biol. **16**: 1089–1091.

Mirkovitch, J., Mirault, M.E., and Laemmli, U.K. 1984. Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell, **39**: 223–232.

Murphy, G.J.P., Lucas, H., Moore, G., and Flavell, R.B. 1992. Sequence analysis of *WIS-2-1A*, a retrotransposon-like element from wheat. Plant Mol. Biol. **20**: 991–995.

Noma, K., Ohtsubo, E., and Ohtsubo, H. 1999. Non-LTR retrotransposons (LIMES) as ubiquitous components of plant genomes. Mol. Gen. Genet. **261**: 71–79.

Panstruga, R., Buschges, R., Piffanelli, P., and Schulze-Lefert, P. 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. Nucleic Acids Res. **26**: 1056–1062.

Payne, P.I. 1987. Genetics of wheat storage proteins and the effect of allelic variation on breadmaking quality. Ann. Rev. Plant Physiol. **38**: 141–153.

Presting, G.G., Malysheva, L., and Schubert, I. 1998. A *Ty3/gypsy* retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. Plant J. **16**: 721–728.

Rampitsch, C., Jordan, M.C., and Cloutier, S. 2000. A matrix attachment region is located upstream from the high-molecular-weight glutenin gene *Bx7* in wheat (*Triticum aestivum* L.). Genome, **43**: 483–486.

Rogowsky, P.M., Liu, J.-Y., Manning, S., Taylor, C., and Langridge, P. 1992. Structural heterogeneity in the R173 family of rye-specific repetitive DNA sequences. Plant Mol. Biol. **20**: 95–101.

SanMiguel, P., Tikhonov, A., Jin, Y.-K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science (Washington, D.C.), **274**: 765–768.

Shewry, P.R., Halford, N.G., and Tatham, A.S. 1992. High molecular weight subunits of wheat glutenin. J. Cereal Sci. **15**: 105–120.

Shewry, P.R., Tatham, A.S., Barro, F., Barcelo, P., and Lazzeri, P. 1996. Biotechnology of breadmaking: unraveling and manipulating the multi-protein gluten complex. Bio/Technology, **13**: 1185–1190.

Shirasu, K., Schulman, A.H., Lahaye, T., and Schulze-Lefert, P. 2000. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res. **10**: 908–915.

Singh, G.B., Kramer, J.A., and Krawetz, S.A. 1997. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic Acids Res. **25**: 1419–1425.

Song, W.-Y., Pi, L.-Y., Bureau, T.E., and Ronald, P.C. 1998. Identification and characterization of 14 transposon-like elements in the noncoding regions of members of the *Xa21* family of disease resistance genes in rice. Mol. Gen. Genet. **258**: 449–456.

Spiker, S., and Thompson, W.R. 1996. Nuclear matrix attachment regions and transgene expression in plants. Plant Physiol. **110**: 15–21.

Spiker, S., Murray, M.G., and Thompson, W.F. 1983. DNase I sensitivity of transcriptionally active genes in intact nuclei and isolated chromatin of plants. Proc. Natl. Acad. Sci. U.S.A. **80**: 815–819.

Vasil, I.K., and Anderson, O.D. 1997. Genetic engineering of wheat gluten. Trends Plant Sci. **2**: 292–297.

Wessler, S.R. 1998. Transposable elements associated with normal plant genes. Physiol. Plant. **103**: 581–586.